

# Estimation for multiplicative models under multinomial sampling

A. Forcina, Dipartimento di Economia, Finanza e Statistica,  
University of Perugia, Italy

April 25, 2017

## Abstract

The models considered in this paper are a special subclass of Relational models which may be appropriate when a collection of independence statements must hold even after probabilities are re-scaled to sum to 1. After reviewing the basic properties of these models and deriving some new ones, two algorithms for computing maximum likelihood estimates are presented. Some new light is also thrown on the underlying geometry.

**Keywords.** Relational models, Curved exponential families, Mixed parametrization, Iterative proportional fitting.

## 1 Introduction

The class of Relational models introduced by Klimova et al. (2012) and developed further by Klimova and Rudas (2016) generalize log-linear models for contingency tables under Poisson or multinomial sampling in at least two interesting directions. First, the cells of the table to which the models can be applied, do not need to be produced by cross classifying a set of discrete random variable. An obvious instance is when attention is restricted to a specific subset. Several interesting examples of applied contexts having these feature may be found in the above papers.

The present paper concentrates on an interesting subclass of relational models: multiplicative models for the vector of probabilities under multinomial sampling. Though these models are meaningful in general, their most common application is when, given the applied context, certain pairs of cells of the contingency table have to satisfy an independence statement and, at the same time, sum to 1. Ordinary log-linear models for a probability vector are unable to achieve this because any independence statement is blurred when probabilities are forced to sum to 1.

It turns out that, to achieve both goals, multiplicative structure and summing to 1, leads to a curved exponential family within the multinomial distribution which has some peculiar features, the main one being that, though the vector of mean parameters of the fitted model is no longer equal to the of vector sufficient statistic divided by the sample size, as in an ordinary log-linear model, still the two vectors are proportional. This result which appeared in Klimova et al. (2012) is the key to computation and interpretation of maximum likelihood estimated (MLEs).

After introducing notations and revising the basic properties of the models in Section 2, Two algorithms for computing MLEs are described in Section 3 where certain insightful geometric properties of these models are also highlighted. Throughout, to make the paper more readable, without the need to browse through the literature and translate to the given context, the most relevant results on multiplicative models are derived directly, though appropriate references are also provided.

## 2 Notations and preliminary results

Let  $V$  denote a discrete random variable with outcomes in  $\mathcal{V} = \{v_1, \dots, v_J\}$ , let  $\pi_j = P(V = v_j)$ ,  $\boldsymbol{\pi}$  denote the vector with elements  $\pi_j$  and assume that  $\boldsymbol{\pi}$  belongs to  $\mathcal{P} = \{\boldsymbol{\pi} : \pi_j > 0, \mathbf{1}'\boldsymbol{\pi} = 1\}$ . Let  $Mn(n, \boldsymbol{\pi})$  denote a multinomial distribution with sample size  $n$  and assume that the objective is to make inference about  $\boldsymbol{\pi}$  based on the vector of observed frequencies  $\mathbf{y} \sim Mn(n, \boldsymbol{\pi})$ .

Let  $\mathbf{X}$  be a  $J \times r$  design matrix whose elements are 0 or 1. Assume that  $\mathbf{X}$  is of full column rank, that it does not contain the row of all 0s and that the unitary vector  $\mathbf{1} = (1, \dots, 1)'$  does not belong to the space spanned by the columns of  $\mathbf{X}$ .

**Definition 1.** *The vector  $\boldsymbol{\pi}$  satisfies a multiplicative model, denoted  $\boldsymbol{\pi} \in \mathcal{M}(\mathbf{X})$ , if*

$$\log \boldsymbol{\pi} = \mathbf{X}\boldsymbol{\beta}, \quad (1)$$

where  $\boldsymbol{\beta}$  is a vector of  $r$  log-linear parameters.

It can be easily verified that (1) holds if and only if, for any  $(J - r) \times J$  matrix  $\mathbf{C}$  of full row rank and such that  $\mathbf{C}\mathbf{X} = \mathbf{0}$

$$\mathbf{C} \log \boldsymbol{\pi} = \mathbf{0}. \quad (2)$$

Multiplicative models under multinomial sampling belong to the class of Relational Models introduced by Klimova et al. (2012).

The following Proposition, which could also be derived from Klimova et al. (2012), indicates that, for any  $\mathcal{M}(\mathbf{X})$ , the constraint matrix  $\mathbf{C}$  can always be written in a kind of canonical form:

**Proposition 1.** *There exists a non singular linear transformation of the constraint matrix  $\mathbf{C}$  such that it can be written in the form*

$$\begin{pmatrix} \mathbf{c}' \\ \mathbf{H} \end{pmatrix}$$

where  $\mathbf{c}'\mathbf{1} = -1$  and  $\mathbf{H}\mathbf{1} = \mathbf{0}$ .

*Proof.* Because  $\mathbf{1}$  does not belong to the space generated by the columns of  $\mathbf{X}$ ,  $\mathbf{C}$  must contain at least a row which does not sum to 0; let  $\mathbf{c}'$  denote the first row of  $\mathbf{C}$  and assume that  $\mathbf{c}'\mathbf{1} = t \neq 0$  (this can be achieved by permuting rows if necessary). Let  $\mathcal{A}$  be the collection of all other rows such that  $\mathbf{c}'_a\mathbf{1} = t_a \neq 0$ ,  $\forall a \in \mathcal{A}$ . To replace  $\mathbf{c}_a$  with  $\mathbf{h}_a = \mathbf{c}_a - \mathbf{c}(t_a/t)$  for all  $a \in \mathcal{A}$  is equivalent to left multiply  $\mathbf{C}$  by a lower triangular matrix which has 1s on the main diagonal and is thus non-singular. The condition that  $\mathbf{c}'\mathbf{1} = -1$  can be achieved by dividing the elements of  $\mathbf{c}$  by a constant different from 0.  $\square$

To see that the multinomial distribution  $Mn(n, \boldsymbol{\pi})$  with  $\boldsymbol{\pi} \in \mathcal{M}(\mathbf{X})$  is a curved exponential family (Theorem 3.2, Klimova et al., 2012), note that the model could be imposed in two steps: (i) first assume that

$$\log \boldsymbol{\pi} = \mathbf{X}\boldsymbol{\theta} - \mathbf{1} \log(\mathbf{1}' \exp(\mathbf{X}\boldsymbol{\theta})), \quad (3)$$

which defines a regular exponential family where the sub-vector of canonical parameters  $\mathbf{H} \log \boldsymbol{\pi}$  is constrained to  $\mathbf{0}$ , let  $\tilde{\boldsymbol{\pi}}$  denote the MLE under this preliminary model; (ii) impose the additional constraint  $\mathbf{c}' \log \boldsymbol{\pi} = 0$ ; this implies that the elements of  $\boldsymbol{\theta}$  have to satisfy the non linear constraint  $\log(\mathbf{1}' \exp(\mathbf{X}\boldsymbol{\theta})) = 0$  which defines an  $r - 1$  dimensional surface.

### 3 Maximum likelihood estimation

Though MLEs for multiplicative models might also be computed by general purpose algorithms like the one described by Evans and Forcina (2013), the approach introduced by Klimova et al. (2012) and studied in detail in Klimova and Rudas (2015) in the context of Relational models, in addition to being, approximately, equally efficient, provides deeper insights into the nature of these models. The objective of this section is to present a new algorithm which could be seen as a dual version of the one in Klimova and Rudas (2015), to describe certain refinements which increase efficiency and help highlight additional relevant features of these models.

Some of the results in the Proposition below could also be derived from Theorem 3.3 in Klimova et al. (2012).

**Proposition 2.** *The likelihood equation can be written in the form*

$$\hat{\gamma} \mathbf{X}' \mathbf{p} - \mathbf{X}' \hat{\boldsymbol{\pi}} = \mathbf{0}, \quad (4)$$

where  $\mathbf{p} = \mathbf{y}/n$  is the vector of cell proportions and  $\hat{\gamma} > 0$ .

*Proof.* Start from the multinomial likelihood with  $\log \boldsymbol{\pi}$  as in (3) and use the Lagrange multiplier  $\alpha$  to account for the additional constraint  $\mathbf{c}' \log \boldsymbol{\pi} = 0$ ; note that, because  $\mathbf{c}' \mathbf{X} = \mathbf{0}'$  and  $\mathbf{c}' \mathbf{1} = -1$ ,  $\mathbf{c}' \log \boldsymbol{\pi} = \log(\mathbf{1}' \exp(\mathbf{X}\boldsymbol{\theta}))$ . To maximize the lagrangian

$$L(\mathbf{y}; \boldsymbol{\pi}, \alpha) = \mathbf{y}' \mathbf{X}\boldsymbol{\theta} - n \log(\mathbf{1}' \exp(\mathbf{X}\boldsymbol{\theta})) - n\alpha(\mathbf{c}' \log \boldsymbol{\pi}),$$

replace  $\mathbf{c}' \log \boldsymbol{\pi}$  with  $\log(\mathbf{1}' \exp(\mathbf{X}\boldsymbol{\theta}))$  and differentiate with respect to  $\boldsymbol{\theta}$  to obtain

$$\mathbf{X}' \mathbf{y} - n(1 + \hat{\alpha}) \mathbf{X}' \hat{\boldsymbol{\pi}} = \mathbf{0}.$$

The result follows by dividing both sides by  $n(1 + \hat{\alpha})$ ; the fact that  $\hat{\gamma} > 0$  follows from Klimova et al. (2012) who call it the adjustment factor.  $\square$

**Remark 1.** *Clearly, if  $\mathbf{c}' \log \tilde{\boldsymbol{\pi}} = 0$ ,  $\hat{\alpha} = 0$  and  $\hat{\gamma} = 1$ . If, instead,  $\mathbf{c}' \log \tilde{\boldsymbol{\pi}} > 0$ , implying that  $\mathbf{1}' \exp(\mathbf{X}\tilde{\boldsymbol{\theta}}) > 1$ , then (4) may be translated into*

$$\mathbf{X}' \tilde{\boldsymbol{\pi}} \hat{\gamma} = \mathbf{X}' \hat{\boldsymbol{\pi}}$$

*which can be satisfied if  $\hat{\gamma} < 1$ . By a similar argument, when  $\mathbf{c}' \log \tilde{\boldsymbol{\pi}} < 0$ ,  $\hat{\gamma}$  must be greater than 1.*

An algorithm for MLE, call it  $M_\pi$ , could consist in solving (4) for a fixed value of  $\gamma$  combined with a search for the value of  $\gamma$  for which  $\mathbf{c}' \log \boldsymbol{\pi} = 0$ . For comparison, the algorithm described in Klimova and Rudas (2015) may be summarized as follows: for a fixed  $\gamma$  use an iterative proportional fitting (IPF) algorithm to solve the problem

$$P_\tau = \begin{cases} \mathbf{X}'\tilde{\boldsymbol{\tau}}(\gamma) = \gamma\mathbf{X}'\mathbf{q}, \\ \mathbf{C} \log \tilde{\boldsymbol{\tau}}(\gamma) = \mathbf{0}, \end{cases} \quad (5)$$

at the same time keep searching for the value  $\hat{\gamma}$  such that the elements of  $\tilde{\boldsymbol{\tau}}(\gamma)$  sum to 1.  $M_\pi$  may be seen as a dual version of  $P_\tau$ : for a fixed  $\gamma$  the estimate  $\tilde{\boldsymbol{\pi}}(\gamma)$  sums to 1 but does not satisfy  $\mathbf{c}' \log \tilde{\boldsymbol{\pi}}(\gamma) = 0$ ; on the other hand,  $\mathbf{c}'\tilde{\boldsymbol{\tau}}(\gamma) = 0$ , but the elements of  $\tilde{\boldsymbol{\tau}}(\gamma)$  do not sum to 1.

**Remark 2.** *The equations in (5) may be interpreted as defining  $\tilde{\boldsymbol{\tau}}(\gamma)$ , the mean vector of an exponential family, by a mixed parametrization (Barndorff-Nielsen, 1978, p. 121-123) which assigns given values to a linear transformation of the mean parameters and constraints the complementary set of canonical parameters to 0. Thus, as an alternative to IPF, (5) may also be solved by the Newton algorithm described in Forcina (2012), Appendix 2. The same algorithm could be used to compute a solution of equation (4) for a given  $\gamma$ : determine the vector of probabilities of a multinomial distribution having the vector of mean parameters  $\gamma\mathbf{X}'\mathbf{p}$  and with the vector of canonical parameters  $\mathbf{H} \log \boldsymbol{\pi} = \mathbf{0}$ .*

### 3.1 Two algorithms for computing the MLE

Let  $f(\gamma) = \mathbf{c}' \log \tilde{\boldsymbol{\pi}}(\gamma)$  and  $g(\gamma) = \log[\mathbf{1}'\tilde{\boldsymbol{\tau}}(\gamma)]$  be the functions that specify the non-homogeneous constraint and the normalization constraint respectively. For a given vector of cell proportions  $\mathbf{p}$ , a value of  $\gamma$  will be called feasible if there exist  $\boldsymbol{\pi} \in \mathcal{P}$  such that (4) holds.

**Lemma 1.** *For all feasible  $\gamma$ ,  $f(\gamma)$  and  $g(\gamma)$  are differentiable and strictly increasing, with:*

$$\begin{aligned} \frac{\partial f(\gamma)}{\partial \gamma} &= \gamma \mathbf{t}' \mathbf{F}(\gamma)^{-1} \mathbf{t}, \\ \frac{\partial g(\gamma)}{\partial \gamma} &= \gamma \mathbf{t}' \mathbf{G}(\gamma)^{-1} \mathbf{t} / [\mathbf{1}'\tilde{\boldsymbol{\tau}}(\gamma)], \end{aligned}$$

where  $\mathbf{t} = \mathbf{X}'\mathbf{p}$ ,  $\mathbf{F}(\gamma) = \mathbf{X}'[\text{diag}(\tilde{\boldsymbol{\pi}}(\gamma)) - \tilde{\boldsymbol{\pi}}(\gamma)\tilde{\boldsymbol{\pi}}(\gamma)']\mathbf{X}$ , and  $\mathbf{G}(\gamma) = \mathbf{X}'\text{diag}(\tilde{\boldsymbol{\tau}}(\gamma))\mathbf{X}$  are positive definite matrices.

*Proof.* See the Appendix. □

The results of Lemma 1, combined with (4) or (5), provide two algorithms for computing the MLE of  $\boldsymbol{\pi} \in \mathcal{M}(\mathbf{X})$ . In both algorithm one can start with  $\gamma = 1$  which is always feasible:

**Algorithm M:**

1. in the  $s$ th step, given  $\gamma^s$ , use a Newton algorithm to compute the vector of probabilities having the set of mean parameters  $\mathbf{X}'\tilde{\boldsymbol{\pi}}(\gamma^s) = \gamma^s\mathbf{X}'\mathbf{p}$  and the vector of canonical parameters  $\mathbf{H} \log \tilde{\boldsymbol{\pi}}(\gamma^s) = \mathbf{0}$ ;
2. use Lemma 1 to update  $\gamma$ :  $\gamma^{s+1} = \gamma^s - f(\gamma^s)/[\gamma^s \mathbf{t}' \mathbf{F}(\gamma^s)^{-1} \mathbf{t}]$ ;

3. iterate until  $f(\gamma^s)$  is not sufficiently close to 0.

**Algorithm P:**

1. in the  $s$ th step, given  $\gamma^s$ , use a Newton algorithm to compute the vector of intensities having the set of mean parameters  $\mathbf{X}'\tilde{\boldsymbol{\tau}}(\gamma^s) = \gamma^s \mathbf{X}'\mathbf{p}$  and the vector of canonical parameters  $\mathbf{C} \log \tilde{\boldsymbol{\tau}}(\gamma^s) = \mathbf{0}$ ;
2. use Lemma 1 to update  $\gamma$ :  $\gamma^{s+1} = \gamma^s - g(\gamma^s)/[\gamma^s \mathbf{t} \mathbf{G}(\gamma^s)^{-1} \mathbf{t}]$ ;
3. iterate until  $g(\gamma^s)$  is not sufficiently close to 0.

Both algorithms seem to be equally fast, usually taking 4 to 8 steps to reach convergence. With  $M_\pi$  it may be wise, initially, to shorten the step length when updating  $\gamma$  to avoid hitting into a non feasible value.

Let  $\tilde{\gamma}_M$  and  $\tilde{\gamma}_P$  be, respectively, the values of  $\gamma$  at convergence of algorithms  $M_\pi$  and  $P_\tau$ .

**Proposition 3.** *At convergence  $\tilde{\gamma}_M = \tilde{\gamma}_P = \hat{\gamma}$  and  $\tilde{\boldsymbol{\pi}}(\hat{\gamma}) = \tilde{\boldsymbol{\tau}}(\hat{\gamma})$ .*

*Proof.* First notice that

$$g(\tilde{\gamma}_P) = \log[\mathbf{1}'\tilde{\boldsymbol{\tau}}(\gamma_P)] = 0 \quad \text{and} \quad g(\tilde{\gamma}_M) = \log[\mathbf{1}'\tilde{\boldsymbol{\pi}}(\gamma_M)] = 0,$$

and thus,  $g(\tilde{\gamma}_P) = g(\tilde{\gamma}_M)$ . Because  $g(\gamma)$  is strictly increasing, this implies  $\tilde{\gamma}_M = \tilde{\gamma}_P$ . When  $\tilde{\gamma}_M = \tilde{\gamma}_P$ , the two vectors  $\tilde{\boldsymbol{\pi}}(\tilde{\gamma}_M)$  and  $\tilde{\boldsymbol{\tau}}(\tilde{\gamma}_P)$  sum to 1, have same mean parameters and satisfy the same set of log-linear constraints, so they must be equal because of the uniqueness of the mixed parametrization.  $\square$

### 3.2 Geometry of MLEs

Recall that a model defined by (3) alone may be interpreted as a log-linear model for a multinomial distribution and its MLEs are determined only by  $\mathbf{t} = \mathbf{X}'\mathbf{p}$ , the vector of sufficient statistics computed on the sample proportions. Let  $\mathbf{s} = \mathbf{X}'\hat{\boldsymbol{\pi}}$ , then all vectors of observations  $\mathbf{y}$  such that the corresponding vector of sample proportions satisfies  $\mathbf{t} = \mathbf{s}$  have the same MLEs.

Equation (4) may be used to characterize the set  $\mathcal{T}(\hat{\boldsymbol{\pi}})$ , with elements  $\mathbf{t} = \mathbf{X}'\mathbf{p}$ , such that all  $\mathbf{t} \in \mathcal{T}(\hat{\boldsymbol{\pi}})$  share the same MLEs under the model  $\mathcal{M}(\mathbf{X})$ . In other words,  $\mathbf{t} \in \mathcal{T}(\hat{\boldsymbol{\pi}})$  if, given  $\hat{\boldsymbol{\pi}}$ , there exists a value  $\hat{\gamma}$  of the adjustment factor such that

$$\mathbf{t} = \mathbf{s}/\hat{\gamma}.$$

Thus, if one divides  $\mathbf{X}'\mathbf{y}$ , the vector of sufficient statistics, by the sample size, the result must lie on the straight line joining the origin with the point  $\mathbf{s}$ ; the actual set of sample points is bounded by the constraint that  $\mathbf{p} \in \mathcal{P}$ . The fact that  $\mathcal{T}(\hat{\boldsymbol{\pi}})$  is linear in  $\mathbf{s}$  is a general property of curved exponential families and follows from (3.3) in Efron (1978); what is interesting here is the special form of this set for multiplicative models

**Example 1.** *For  $J = 4$  consider the design matrix*

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix}.$$

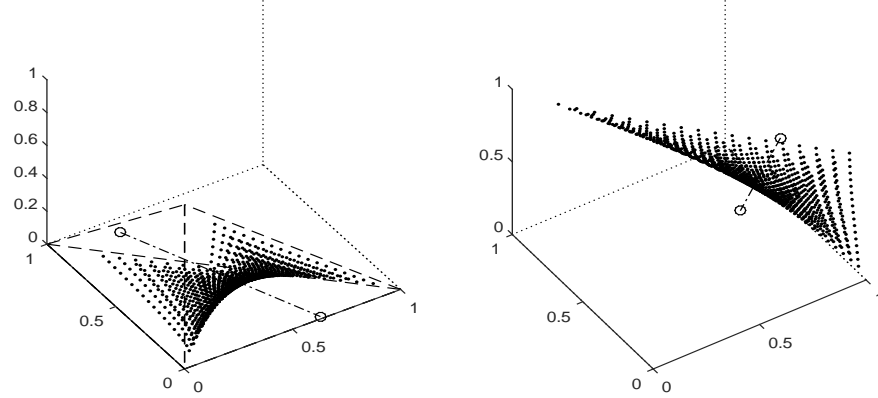


Figure 1: For each vector of sample proportions belonging to a 0.05 grid within  $\mathcal{P}$ , the first three coordinates of  $\hat{\pi}$  are plotted on the left side; the corresponding values of  $\mathbf{X}'\hat{\pi}$  are plotted on the right hand side. For a specific value of  $\hat{\pi}$ , the linear set  $\mathcal{G}(\hat{\pi})$  is plotted on the left side by a dash-dot line with circles marking end points. The corresponding set  $\mathbf{X}'\hat{\pi}/\hat{\gamma}$  for all feasible  $\hat{\gamma}$  is plotted on the right side.

Because  $\pi \in \mathcal{P}$ , the fourth dimension is redundant and sample proportions can be plotted in three dimensions. For each  $\mathbf{p}$  on a grid within  $\mathcal{P}$ , the resulting  $\hat{\pi}$  are plotted in Figure 1 (left) together with the corresponding value of  $\mathbf{s} = \mathbf{X}'\hat{\pi}$  (right). Each set of points outline a corresponding two dimensional surfaces.

Notice that the pair  $(\hat{\pi}, \hat{\gamma})$  determine a single  $\mathbf{t}$ ; clearly, all the vectors of sample proportions  $\mathbf{p}$  such that  $\mathbf{X}'\mathbf{p} = \mathbf{t}$  share the same MLEs under (3) and the same value of the log-likelihood. There are two special points worth mentioning within  $\mathcal{T}(\hat{\pi})$ :

1. the one with  $\mathbf{t} = \mathbf{s}$  and  $\hat{\gamma} = 1$ ; a very special instance is when  $\mathbf{p} = \hat{\pi}$ ;
2. the one corresponding to the smallest feasible adjustment factor, say,  $\hat{\gamma}_L$ , which lies on the boundary of the sample space in the direction opposite to the origin

The following Proposition shows that, contrary to intuition, the vector  $\mathbf{t}$  giving the largest likelihood is the one corresponding to  $\hat{\gamma}_L$  rather than to  $\hat{\gamma} = 1$ .

**Proposition 4.** *Among the vectors  $\mathbf{t} \in \mathcal{T}(\hat{\pi})$ , the one associated with  $\hat{\gamma}_L$  has the largest likelihood.*

*Proof.* See the Appendix. □

Because  $\hat{\gamma}$  is smallest when  $\hat{\alpha}$  is largest, Remark 1 implies that, when  $\hat{\gamma} = \hat{\gamma}_L$ ,  $\mathbf{c}' \log \hat{\pi}$  is negative.

## Acknowledgments

The author would like to thank A. Klimova and T. Rudas for sharing ideas concerning Relational models and for several very enlightening discussions.

## Appendix

### Proof of Lemma 1

For simplicity, in the following write  $\boldsymbol{\tau}$  for  $\hat{\boldsymbol{\tau}}(\gamma)$  and similarly  $\boldsymbol{\pi}$  for  $\hat{\boldsymbol{\pi}}(\gamma)$ . To differentiate  $g(\gamma)$ , let  $\boldsymbol{\mu} = \mathbf{X}'\boldsymbol{\tau} = \gamma\mathbf{X}'\mathbf{p}$ , and recall that  $\boldsymbol{\tau} = \exp(\mathbf{X}\boldsymbol{\theta})$  then, by the chain rule,

$$\frac{\partial g(\gamma)}{\partial \gamma} = \frac{\partial g(\gamma)}{\partial \boldsymbol{\theta}'} \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\mu}'} \frac{\partial \boldsymbol{\mu}}{\partial \gamma} = (\mathbf{1}'\boldsymbol{\tau})^{-1} \boldsymbol{\tau}' \mathbf{X} \left( \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}'} \right)^{-1} \mathbf{X}' \mathbf{p}$$

Replace  $\boldsymbol{\tau}' \mathbf{X}$  with  $\gamma \mathbf{p}' \mathbf{X}$ ; because  $\boldsymbol{\mu} = \mathbf{X}' \exp(\mathbf{X}\boldsymbol{\theta})$ ,

$$\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}'} = \mathbf{X}' \text{diag}(\boldsymbol{\tau}) \mathbf{X}.$$

To differentiate  $f(\gamma)$  recall (3) and note that (4) implies  $\boldsymbol{\mu} = \mathbf{X}'\boldsymbol{\pi} = \gamma\mathbf{X}'\mathbf{p}$ . Because, by assumption,  $\mathbf{c}'\mathbf{X} = 0$  and  $\mathbf{c}'\mathbf{1} = -1$ ,  $f(\gamma) = \log[\mathbf{1}' \exp(\mathbf{X}\boldsymbol{\theta})]$ , by the chain rule

$$\frac{\partial f(\gamma)}{\partial \gamma} = \frac{\partial \log[\mathbf{1}' \exp(\mathbf{X}\boldsymbol{\theta})]}{\partial \gamma} = \frac{\exp(\mathbf{X}\boldsymbol{\theta})' \mathbf{X}}{\mathbf{1}' \exp(\mathbf{X}\boldsymbol{\theta})} \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\mu}'} \mathbf{X}' \mathbf{p}.$$

The result follows because, by construction,  $\exp(\mathbf{X}\boldsymbol{\theta})' \mathbf{X} / [\mathbf{1}' \exp(\mathbf{X}\boldsymbol{\theta})] = \boldsymbol{\pi}' \mathbf{X} = \gamma \mathbf{p}' \mathbf{X}$  and

$$\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\mu}'} = \left( \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}'} \right)^{-1} = \mathbf{X}' \frac{\partial \boldsymbol{\pi}}{\partial (\mathbf{X}\boldsymbol{\theta})'} \mathbf{X}.$$

### Proof of Proposition 4

Recall that, at convergence,  $\hat{\boldsymbol{\pi}}$  is equal to  $\hat{\boldsymbol{\tau}}$ , the solution to  $P_{\boldsymbol{\tau}}$ . Let  $\mathbf{X}^- = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ ,  $\mathbf{C}^- = \mathbf{C}'(\mathbf{C}\mathbf{C}')^{-1}$  and note that the matrix  $\mathbf{G} = \begin{pmatrix} \mathbf{X} & \mathbf{C}^- \end{pmatrix}$  is  $J \times J$  and has an explicit inverse so that

$$\mathbf{G}^{-1} \log \hat{\boldsymbol{\tau}} = \begin{pmatrix} \mathbf{X}^- \\ \mathbf{C} \end{pmatrix} \log \hat{\boldsymbol{\tau}}.$$

Because  $\hat{\boldsymbol{\tau}}$  satisfies the second equation in  $P$ ,  $\mathbf{C} \log \hat{\boldsymbol{\tau}} = \mathbf{0}$ , it follows that

$$LL(\mathbf{y}) = \mathbf{y}' \log \hat{\boldsymbol{\tau}} = \mathbf{y}' \mathbf{G} \mathbf{G}^{-1} \log \hat{\boldsymbol{\tau}} = n(\mathbf{X}'\mathbf{p})' \mathbf{X}^- \log \hat{\boldsymbol{\tau}} = n\mathbf{s}' \mathbf{X}^- \log \hat{\boldsymbol{\tau}} / \hat{\gamma};$$

. thus the maximum is achieved when  $\hat{\gamma}$  equals the smallest possible value in  $\mathcal{G}(\hat{\boldsymbol{\pi}})$

## References

- Barndorff-Nielsen, O. E. (1978). *Information and exponential families*. Wiley, New York.
- Efron, B. (1978). The geometry of exponential families. *The Annals of Statistics*, 6:362–376.
- Evans, R. J. and Forcina, A. (2013). Two algorithms for fitting constrained marginal models. *Comput. Statist. Data Anal.*, 66:1–7.
- Forcina, A. (2012). Smoothness of conditional independence models for discrete data. *J. Multivariate Anal.*, 106:49–56.

- Klimova, A. and Rudas, T. (2015). Iterative Scaling in Curved Exponential Families. *Scand. J. Statist.*, 42:832–847.
- Klimova, A. and Rudas, T. (2016). On the closure of relational models. *J. Multivariate Anal.*, 143:440–452.
- Klimova, A., Rudas, T., and Dobra, A. (2012). Relational models for contingency tables. *J. Multivariate Anal.*, 104:159–173.